PB90-125873

Comparison of Robust Bayes and Classical Estimators for Regional Lake Models of Fish Response to Acidification

Duke Univ., Durham, NC

Prepared for:

Corvallis Environmental Research Lab., OR

1988

U.S. Department of Commerce
National Technical Information Service
NTIS

ERL-COR-791JAP

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA/600/J-88/433	2.
3. RI PB90-125873	
4. TITLE AND SUBTITLE A comparison of robust Bayes and classical estimators for regional lake models of fish response to acidification	5. REPORT DATE
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Kenneth H. Reckhow	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Forestry & Environmental Studies, Duke University	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS US Environmental Protection Agency Environmental Research Laboratory 200 SW 35th Street Corvallis, OR 97333	13. TYPE OF REPORT AND PERIOD COVERED Journal Article
	14. SPONSORING AGENCY CODE EPA/600/2

15. SUPPLEMENTARY NOTES

Water Resources Research 24(7):1061-1058. July 1988.

16. ABSTRACT

Empirical models of fish response to lake acidification were recently fit to a large historical data set from the Adirondack region of the United States using classical and Bayesian methods. The models may be used to predict species presence/absence for brook trout and lake trout as a function of acid-precipitation-related water chemistry, using a logistic function. To evaluate the effectiveness of the models in the prediction of presence/absence due to regional lake acidification new data sets were used for cross validation of the candidate models. Based on this evaluation, the robust Bayes models, which are based on a compromise estimator between Bayes and empirical Bayes, were found to be the best predictors of species presence/absence in lakes.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL INFORMATION SERVICE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS (This Report) Unclassified	21. NO. OF PAGES 10
	20. SECURITY CLASS (This page) Unclassified	22. PRICE

EPA Form 2220-1 (9-73)

EPA/600/J-88/433

WATER RESOURCES RESEARCH, VOL. 24, NO. 7, PAGES 1061–1068, JULY 1988

A Comparison of Robust Bayes and Classical Estimators for Regional Lake Models of Fish Response to Acidification

KENNETH H. RECKHOW

School of Forestry and Environmental Studies, Duke University, Durham, North Carolina

Empirical models of fish response to lake acidification were recently fit to a large historical data set from the Adirondack region of the United States using classical and Bayesian methods. The models may be used to predict species presence/absence for brook trout and lake trout as a function of acid-precipitation-related water chemistry, using a logistic function. To evaluate the effectiveness of the models in the prediction of presence/absence due to regional lake acidification, new data sets were used for cross validation of the candidate models. Based on this evaluation, the robust Bayes models, which are based on a compromise estimator between Bayes and empirical Bayes, were found to be the best predictors of species presence/absence in lakes.

INTRODUCTION

The conventional approach for the determination of parameters for mechanistic water quality simulation models is for the modeler to use his judgment in selecting parameters that are consistent with any available data as well as with tabulations of accepted coefficient values [e.g., *Zison et al.*, 1978]. Parameters are often selected in a sequential manner (perhaps based on a sensitivity analysis), and the model is finally judged adequate based on a visual comparison of predictions and observations. Formal mathematical optimization is not usually involved.

Parameter estimation in empirical water quality models has traditionally been undertaken using classical optimization methods such as maximum likelihood or least squares. Judgment is typically involved in the specification of the model but not in the actual estimation algorithm.

The difference in approach between these two categories of models has occurred for a few reasons. In some cases, mechanistic modelers have believed that the model equations are theoretically correct, and the model parameters are physically measurable quantities that are simply measured in the field and inserted into the model. In other cases, the mechanistic models, with large numbers of parameters, were not identifiable without imposing constraints on most of the parameters (and effectively estimating these parameters using expert judgment and literature tabulations). In contrast, empirical models, with one or only a few parameters, are identifiable from the available data, and the empirical modeler believes that the model is more credible if an optimality criterion is used to estimate the parameters.

In many situations, where available data and expert judgment permit, it may be to the advantage of the water quality modeler to use both site specific data and expert judgment in the estimation of model parameters based on estimators with optimal properties. For example, Bayesian inference provides the framework for the pooling of judgmental probabilities and sample information; estimators for model parameters based on the posterior distribution have optimal characteristics in a Bayesian sense, and also in a classical sense conditional on the truth of the prior (judgmental) probability [*Berger*, 1984].

Copyright 1988 by the American Geophysical Union.

Paper number 7W5001.
0043-1397/88/007W-5001$05.00

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

One problem with Bayesian inference is in the determination of the judgmental probability model that correctly represents expert opinion. It is well known [*Kahneman et al.*, 1982] that quantification of probabilistic judgment is a difficult task, particularly for those who are unfamiliar with expressing their opinions in this way. While a Bayesian might argue that an inadequate prior probability model can be dealt with if more time is allocated to elicitation of the prior or if a second-stage prior is created to reflect uncertainty in the first-stage prior, time limitations and complicated estimators may favor other approaches.

Accordingly, a robust Bayesian analysis [*Berger*, 1982] is considered here to guard against misspecification of the prior. In brief, the robust Bayes estimator provides robustness against the prior by, in this case, selecting between a Bayes estimator and a James-Stein estimator according to the closeness between the prior and the data.

The problem of interest is the development and evaluation of empirical models of fish response to lake acidification. Experimental evidence indicates that acid precipitation is a likely cause of the absence of fish populations in selected North American lakes. To support this scientific information and to provide a firm basis for policy making, observational evidence is necessary. Based on the effectiveness of maximum likelihood models [*Reckhow et al.*, 1987], brook trout and lake trout were selected for development of robust Bayes models. The primary chemistry variables considered were *p*H and calcium (Ca), both of which have been found to be important in experimental and observational studies of fish survival [*Baker and Schofield*, 1982; *Brown*, 1982; *Baker*, 1985].

The work summarized herein was undertaken to develop and evaluate empirical models of fish response to lake acidification. In the following sections the methods of maximum likelihood, Bayesian analysis, empirical Bayes, and robust Bayes are discussed. Maximum likelihood and robust Bayes methods are used to estimate parameters for a presence/absence fish response model. Additional information concerning the data analysis and maximum likelihood model development is presented elsewhere [*Reckhow et al.*, 1987].

METHODS OF PARAMETER ESTIMATION

Three basic methods of parameter estimation were used at various stages in the development and evaluation of empirical models of fish response. These methods are maximum likelihood logistic regression analysis, Bayesian analysis, and em-

pirical Bayes methods. The model used in this study is the cumulative logistic probability distribution function:

$$Pr(P)_i = 1/(1 + \exp - (\beta_k x_{ik})) \quad (1)$$

where x_{ik} are observable exogenous variables (x_1 is a column of ones), β_k are unobservable parameters (β_1 is an intercept term), and $Pr(P)_i$ is the probability that case i has the fish species present. The logistic model has domain 0–1, which leads to the interpretation of the response variable as a probability.

The parameters of the model can be estimated using maximum likelihood estimation; details of the estimation procedure are given by *Pindyck and Rubinfeld* [1981]. The assumption inherent in this procedure is that the probability of an event (fish species presence or absence), conditional on x_k, is correctly specified by (1). Maximum likelihood estimation of the model parameters results in consistent and asymptotically (large sample) efficient and normal estimators (Pindyck and Rubinfeld). Thus t-statistics can be applied to test the significance of individual parameters.

Another frequently used procedure for binary (or multiple) choice model parameter estimation is discriminant analysis [*Green*, 1978]. For a categorical dependent variable, with equal predictor variable covariance matrices within each of the groups, a linear discriminant function may be estimated by maximizing the ratio of the between groups variation to the within groups variation. A vector of weights results, and these weights are the parameters in the discriminant function. If it is assumed that the predictor variables are distributed within each group according to a known multivariate normal distribution and the groups have equal covariance matrices, then Bayes Theorem may be used to estimate the posterior probability of group classification with the logistic model.

Although most Monte Carlo studies have shown the general superiority of maximum likelihood logistic regression over discriminant analysis for dichotomous response variables [*Efron*, 1975; *Press and Wilson*, 1978; *Maddala*, 1983], there are also indications of excellent performance by discriminant function estimators with real data sets [*Halperin et al.*, 1971; *Rolph et al.*, 1979]. In fact, under multivariate normality the discriminant function estimators are the true maximum likelihood estimators and thus will be assymptotically more efficient than the conditional maximum likelihood logistic regression estimators [*Maddala*, 1983]. Accordingly, we used discriminant function estimators because of their equivalence to the linear regression estimators which result from the Bayesian analysis.

Bayesian inference results from the application of Bayes Theorem, which provides a logical framework for pooling information from more than one source. Bayes Theorem becomes controversial when subjective information is expressed as a (prior) probability model and then is formally combined with the data-based likelihood function to yield a posterior probability distribution. In the case of the fish response model, the prior distribution reflects expert judgment concerning the model parameters (prior to viewing the data), the likelihood function characterizes the parameters based on the actual data on water chemistry and fish species response, and the posterior distribution represents a pooling of these two sources of information to describe the model parameters, conditional on the observed sample.

Empirical Bayes (EB) methods also use Bayes Theorem, but otherwise, they are more classical (or frequentist) than Bayesian (see *Morris* [1983], including comments). In the typical parametric empirical Bayes problem we wish to simultaneously estimate parameters $\mu_1, \cdots, \mu_p$ (e.g., p means). Data exist for each of the p parameters, and a model is proposed for the data; thus a maximum likelihood estimate could be obtained. However, a second (prior) model is assumed to correctly describe an underlying relationship among the μ_p and Bayes Theorem is used to define the estimators for the posterior parameters. Then under the EB approach the parameters of the prior model are estimated (often using maximum likelihood) from the marginal distribution for the data. In contrast, a Bayesian would likely place a "second-stage prior" on the unknown prior parameters and use Bayes Theorem to pool all information. Empirical Bayes methods are closely linked to the James-Stein estimator [*Berger*, 1982], which is a nonlinear biased estimator that shrinks the maximum likelihood estimator toward a point (often the origin or the grand mean). The James-Stein estimator dominates the maximum likelihood estimator under normality, for a classical loss function of aggregate (over p) squared error, when $p = 3$ or more. This makes EB estimators appealing in a classical sense, when the number of parameters to be simultaneously estimated is large. For instructive treatments of empirical Bayes, James-Stein, and Bayesian analyses in hydrology, see Kuczera [1982], *Landwehr et al.* [1984], and *Marin* [1986], respectively.

When the number of parameters to be estimated is small, a pragmatic (as opposed to a pure Bayesian or pure frequentist) modeler may prefer a Bayesian approach over EB for a normal linear model, since the so-called "Stein-effect" (dominance over maximum likelihood) does not hold for small p. Further, until *Morris* [1983], only asymptotic confidence intervals were determined for empirical Bayes analyses. Unfortunately, there is no guarantee that the prior is correctly specified in a Bayesian analysis; misspecification of the prior can result in a Bayesian estimator with expected loss greater than that for the maximum likelihood estimator over regions of the parameter space. A Bayesian might object to this statement using a classical error criterion to evaluate a point estimate based on the posterior density function. However, it seems reasonable that there should be merits to an estimator that is based on a Bayesian posterior distribution yet is good in a classical risk sense, particularly if one is quite uncertain about the parameter value.

In response to concerns about EB methods for small p and Bayes methods under misspecification of the prior, *Berger* [1982, 1984] has proposed robust Bayes estimators for EB problems. These estimators are related to the "limited translation" estimators of *Efron and Morris* [1971, 1972], which are intended to guard against shrinkage toward an inappropriate prior subpopulation model. Berger's robust Bayes (RB) estimators select between Bayes and empirical Bayes, depending upon the closeness of the Bayes prior to the data.

For the fish response models an expert (Dr. Joan Baker) in fish response to acidification was interviewed to elicit a prior for the model parameters. Since the expression of knowledge as a prior probability is not a routine task for Dr. Baker, it seems advisable to guard against misspecification of a judgmental prior. This suggests that an empirical Bayes analysis may be preferable, since there are two fish species (brook trout and lake trout), and an EB prior (as opposed to a Bayesian judgmental prior) could be constructed based on the data. However, since this is a small-p empirical Bayes problem ($p = 2$ species), the Stein-effect cannot be counted on to guarantee improvement in frequentist risk over the maximum

TABLE 1. Descriptive Statistics for Brook Trout Data

Variable	N	Mean	Standard Deviation	Minimum	Maximum
			Absence		
*p*H	37	5.534	1.014	4.505	7.990
ln(calcium)	30	4.723	0.613	3.761	6.646
ln(aluminum	14	5.336	0.543	3.912	6.040
ln(alkalinity)	47	5.169	0.464	4.620	7.098
ln(surface area)	48	−1.518	1.804	−4.828	3.340
			Presence		
*p*H	40	6.398	0.726	4.820	7.503
ln(calcium)	35	5.039	0.465	4.174	5.956
ln(aluminum)	7	4.408	1.504	2.303	5.991
ln(alkalinity)	52	5.402	0.361	4.812	6.196
ln(surface area)	54	−1.592	1.685	−4.423	3.061

Calcium and alkalinity are in ueq/L, aluminum is in ug/L, and surface area is in km^2. Alkalinity is shifted by +150 to create strictly positive values.

likelihood estimator. Thus the Bayes approach does seem appropriate. Accordingly, Berger's robust Bayes estimators are applied here to achieve some gains from a Bayesian approach in an EB problem, while restricting the risk associated with misspecification of the prior.

Several practical problems limit the inferences that may be drawn from empirical models of fish response with observational data. In brief, the fish response models are certainly misspecified, errors of observation are nonnegligible, and the predictor variables are correlated. General solutions have been proposed in the literature for all three of these problems. Usually the best solution, and often the most difficult solution to carry out, is to gather more information. Better scientific knowledge, data collected under experimental control of exogenous variables, or data on error terms, can be used to help solve each problem. Methods of parameter estimation based on Bayes Theorem also offer the attractive feature of addition of collateral information, and it is for that reason that they are applied and presented here.

Given the consequences of these problems on the model parameters, it is inappropriate to infer cause-effect from the "significance" of estimated parameters in these empirical fish response models. It is reasonable to demand models that are not inconsistent with theory; however, it is unreasonable to adopt a new theory to explain particular parameter estimates for a logistic fish response model. The models are simply summarizing data associations, and it is likely that there are conditions other than strict cause-effect that determine parameter estimates and associated error terms.

MAXIMUM LIKELIHOOD MODELS OF FISH RESPONSE

Previous studies [e.g., *Beggs et al.*, 1985; *Wright and Snekvik*, 1977; *Brown and Sadler*, 1981; *Muniz et al.*, 1984; *Sevaldrud and Muniz*, 1980] involving cross-sectional empirical analyses of fish response to lake acidification provide useful background information and confirm the many experimental studies demonstrating the importance of *p*H and calcium in fish survival.

In this study a large cross-sectional data set (Fish information Network, or FIN) based on historical information on lake chemistry and fish surveys in the Adirondacks of New York State was obtained from *Baker and Harvey* [1984]. These data were originally acquired by many different investigators for a variety of purposes using a variety of analytical methods. Statistical, ecological, and policy considerations resulted in certain constraints imposed on the data set for the purpose of empirical model development. Thus analyses reported herein concern mean values, for June–September data, taken at 0- to 1-m depth, in lakes not recently (within 5 years of water chemistry data used) stocked, for which fish species presence was confirmed at some point in the past. For each fish species, only gill net survey data were used, and the lake chemistry data used were confined to a single analytical method. These constraints reduced the size of the data set but helped to eliminate nuisance variation.

Descriptive statistics for the brook trout sample are presented in Table 1. All concentration variables (and surface area) were log transformed. Kolmogorov-Smirnov tests of fit were deemed inconclusive in favoring a particular probability model, so the lognormal model was adopted as it has often been found to adequately describe marginal distributions of water quality data. Sample size was not constant across variables, reflecting the incomplete nature of the historical FIN data. For the brook trout sample, Table 1 indicates that the mean "absence" lake has lower *p*H, lower calcium, higher aluminum, lower alkalinity, and less surface area than does the mean "presence" lake. These are all consistent with theory.

Given the small sample size for aluminum and the redundancy between alkalinity and the other variables (*p*H and calcium), it was decided to specify the fish response models as a function of *p*H and in calcium. Using maximum likelihood logistic regression, the following model was fitted for brook trout:

$$Pr(P) = 1/[1 + \exp\{2.303 - 2.493pH + 2.615 \ln (Ca)\}] \quad (2)$$
$$(2.783)\ (0.8416)\ (1.304)$$

where $Pr(P)$ is the response variable "probability of presence," and the numbers in parentheses below the model parameters are the standard errors. This model was fitted to a 46-lake subset (all lakes with both *p*H and calcium data) of FIN which contained 23 observed absences and 23 observed presences. An optimistic measure of the effectiveness of (2) for classification is calculated as the percentage of lakes in the 46-lake subset that is correctly classified by the model (using $Pr(P) = 0.5$ as the presence/absence cutoff). This measure is optimistic because the same data set is used to both estimate parameters and to evaluate the fit; in a later section, cross validation is used to evaluate the models using a different data

TABLE 2. Descriptive Statistics for Lake Trout Data

Variable	N	Mean	Standard Deviation	Minimum	Maximum
			Absence		
*p*H	17	5.798	0.939	4.725	7.503
ln(calcium)	11	4.759	0.462	4.172	5.892
ln(aluminum)	6	4.940	0.216	4.700	5.347
ln(alkalinity)	17	5.172	0.375	4.852	6.194
ln(surface area)	17	−0.450	1.229	−2.560	1.763
			Presence		
*p*H	15	6.443	0.691	5.107	7.400
ln(calcium)	9	5.051	0.454	4.443	5.956
ln(aluminum)	4	4.704	1.180	3.689	6.273
ln(alkalinity)	15	5.346	0.326	4.965	6.137
ln(surface area)	15	−0.530	1.208	−2.254	1.744

Calcium and alkalinity are in ueq/L, aluminum is in ug/L, and surface area is in km^2. Alkalinity is shifted by +150 to create strictly positive values.

set. According to this statistic, 19.6% (9/46) of the lakes are incorrectly classified using the model. Unbiased measures of classification error are discussed by *Efron* [1986] and applied to this model by *Reckhow et al.* [1987].

Similar results were found in an analysis of lake trout data in FIN. Table 2 presents statistics for the lake trout sample. Based on considerations of sample size, redundancy, and the need to match the brook trout specification (for EB analysis), the lake trout model was specified as a function of *p*H and ln calcium.

Using maximum likelihood logistic regression, the following model was fitted for lake trout:

$$Pr(P) = 1/[1 + \exp\{2.717 - 2.442pH + 2.585 \ln (Ca)\}] \quad (3)$$
$$(5.767)\ (1.421)\ \ (2.457)$$

The model was fitted to a 20-lake subset (all lakes with both *p*H and calcium data) of FIN which contained 11 observed absences and 9 observed presences. Based on the same optimistic error criterion as applied to the brook trout model, 35% (7/20) of the lakes are incorrectly classified using the model.

Note that the *p*H and ln (Ca) parameters have opposite signs in both (2) and (3). This result is contrary to current scientific understanding which associates both low *p*H and low calcium levels with fish population loss. *Reckhow et al.* [1987] observe that the opposite signs are due to the strong correlation (multicollinearity) between *p*H and ln (Ca). Models for either species fit to only a single predictor variable (*p*H or ln (Ca)) yield the same sign (negative) for each variable [*Reckhow et al.*, 1985]. When both variables are simultaneously in the model, one frequent consequence of multicollinearity is counterintuitive signs on parameters; another consequence is inflated parameter errors. These problems were discussed by *Reckhow et al.* [1987], and they evaluate alternative estimation approaches (e.g., principal components regression) to create a model that has signs consistent with theory. This issue is discussed again below in conjunction with the elicited prior probability model and the robust Bayes estimates.

ROBUST BAYES MODELS

To apply Berger's RB estimator to each fish species, it is first necessary to construct a prior probability model for the three parameters in (2) and (3). Since scientific experts in fish response to acidification think in terms of the variables (*p*H, calcium, and presence/absence) rather than in terms of the model parameters, a predictive distribution elicitation approach [*Winkler*, 1977; *Winkler et al.*, 1978] was used to determine the prior. For this procedure the expert (Dr. Joan Baker) was given a set of exogenous variables and then asked to give her estimate of the median response. Thus a typical question was: "Given 100 lakes in the Adirondacks that have supported brook (lake) trout populations in the past, and if all 100 lakes have *p*H = 5.6 and calcium concentration = 130 μeq/L, what number do you now expect continue to support the brook (lake) trout population?" This question was repeated 20 times with a variety of *p*H-calcium pairs to yield 20 predicted responses. Twenty was chosen to provide some redundancy to improve characterization of the prior yet not burden the expert with time-consuming questions. The *p*H-calcium pairs were not randomly selected but rather were chosen to closely resemble the data (**X**) matrix for the FIN data subsets. This is important in later calculations.

The expected response provided by Dr. Baker does not provide a crucial measure of error. Thus it was assumed that the errors in the conditional response were approximately normally distributed, and additional questions were posed to Dr. Baker to determine fractiles of the predictive distribution, conditional on **X** (*p*H and calcium). A typical question was: "For *p*H = 5.1 and calcium = 90 μeq/L, you estimated that 55 lakes supported brook trout populations. If the odds are 3:1 that the number of lakes (of 100 total lakes) supporting brook trout is greater than a particular value, what is that value?" This question yields the 25th%, and other similar questions provide other percentiles. These fractiles were assessed for six conditional y distributions, producing six estimates for standard error which are conditional on an assumed known underlying variance (estimated from the data). It was decided to be conservative and select the largest value (0.137) for this error variance term, since it is possible that truncation may have affected some of the fractiles. This concern arises because the fractiles are constrained between 0 and 100 by Dr. Baker, but they are not constrained for the linear data-based models.

As applied here, Berger's RB estimator requires independent calculations to select between Bayes and EB estimators for each parameter. To accomplish this, the linear orthonormal model is used to eliminate parameter correlations and allow

independent estimation of each parameter. Fortunately, the ordinary least squares (OLS) regression model can be reexpressed in orthonormal form using singular value decomposition [*Mandel*, 1982], and OLS regression can be used to estimate discriminant function parameters [*Rolph et al.*, 1979], which in turn can be inserted directly into the exponential argument of the logistic function in (1). Thus the models are initially expressed linearly as

$$y = \alpha_1 + \alpha_2 p\mathrm{H} + \alpha_3 \ln(\mathrm{Ca}) + \varepsilon \tag{4}$$

where y is 0, 1 for absence or presence, α_1 through α_3 are the model parameters, and ε is an error term. In matrix terms this model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{5}$$

where the lowercase vectors y, $\boldsymbol{\alpha}$, and $\boldsymbol{\varepsilon}$ are $N \times 1$, 3×1, and $N \times 1$, respectively, and the uppercase data matrix **X** is $N \times 3$. Using singular value decomposition, (5) may be reexpressed as

$$\mathbf{y} = \mathbf{UHV}'\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{6}$$

where the columns of **U** are the eigenvectors of $\mathbf{XX}'$, the rows of $\mathbf{V}'$ are the eigenvectors of $\mathbf{X}'\mathbf{X}$, and the diagonal elements of diagonal matrix **H** are the square roots of the nonzero eigenvalues of $\mathbf{X}'\mathbf{X}$. The parameters are estimated for the reexpression:

$$\mathbf{y} = \mathbf{U}\mathbf{a} + \boldsymbol{\varepsilon} \tag{7}$$

where $\mathbf{a} = \mathbf{HV}'\boldsymbol{\alpha}$. This reparameterization is used because it results in uncorrelated parameters ($\mathbf{a}$) that may be estimated independently, using:

$$\hat{\mathbf{a}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{y} = \mathbf{U}'\mathbf{y} \tag{8}$$

since $\mathbf{U}'\mathbf{U} = \mathbf{I}$.

Thus the approach employed involved reexpression of each data set (FIN brook trout, FIN lake trout, and the data matrix from Dr. Baker's elicitation) in the coordinate system defined by the eigenstructure of the marginal (brook and lake trout) data matrix. The combined (brook and lake trout) data matrix was selected for determination of eigenvectors and eigenvalues since it contained all observational data. This new orthogonal coordinate system (based on the eigenvectors) was then imposed on the individual data sets (the sample brook trout and lake trout data and the prior pseudodata for brook trout and lake trout).

If **X** for FIN brook trout, FIN lake trout, and Dr. Baker's priors were all identical, then each would have the same eigenstructure. This would mean that the same set of coordinates (in **U**) would result in uncorrelated parameters in all data sets. In this study the X matrices were similar but not identical. For example, the data matrix used in the prior elicitation was intentionally structured to resemble the observational data matrix but was not identical to that matrix; the difference resulted from the fact that time restricted the prior elicitation to only 20 pseudoobservations.

Thus model parameters estimated using the individual data sets in the orthogonal coordinate system were not always uncorrelated, as they were (by definition) in the combined data set. However, all correlations were small (less than 0.10), so it was decided that the effect of ignoring these correlations through independent estimation of each parameter was minor and could be neglected. This effect is somewhat similar to the implied spherical (uncorrelated) prior at the origin imposed on the parameters when ridge regression is applied. However, in the model discussed in this paper the correlations are small; whereas, ridge regression is often applied in problems involving high parameter correlations (due to multicollinearity).

Parameters ($\mathbf{a}_{OLS}$) were estimated for each separate data set (prior and sample, brook and lake trout) using OLS, and these were converted to discriminant function parameters ($\mathbf{a}_{DF}$) using *Rolph et al.*'s [1979] simple reexpressions:

$$(a_1)_{DF} = \log(n_1/n_2) + (n/SS_e)[(a_1)_{OLS} - 0.5) + n(n_1^{-1} - n_2^{-1})/2 \tag{9}$$

$$(a_{2,3})_{DF} = (n/SS_e)(a_{2,3})_{OLS} \tag{10}$$

with parameter variances (v_i) approximated by

$$(v_i)_{DF} = (n/SS_e)^2(v_i)_{OLS} \tag{11}$$

where n_1 and n_2 refer to the number of observations in the absence and presence categories, $n = n_1 + n_2$, and SS_e is the residual sum of squares. Berger's RB estimator was determined for the discriminant function parameters in **U** space and then converted back to the original metric using the equations above.

Berger's robust Bayes estimator is based on a comparison of estimators according to classical risk criteria. Specifically, Berger's RB estimator involves comparison of candidate estimators with the Bayes estimator at one extreme and the maximum likelihood estimator at the other extreme. Selection of the RB estimator requires a decision on the trade-off between the additional risk of a robust Bayes estimator when the Bayes estimator dominates and the possible additional risk of the RB estimator in comparison to the maximum likelihood estimator. In the equal variance case, when $p = 3$ or more, Berger's estimator is minimax. However, for $p = 2$ the maximum likelihood estimator is minimax.

For this problem it is decided that we will accept an estimator that may be as much as 20% higher in risk than the maximum likelihood estimator in certain regions of parameter space, while sacrificing only about 20% of the gain in risk from Bayesian analysis. This is a pessimistic expression of the risk; in an optimistic vein, if the Bayes prior is the true prior, we expect lower risk than that from maximum likelihood, and if the Bayes information is misspecified, we expect lower risk than that from the Bayesian analysis.

For $p = 2$, Berger's RB estimator is

$$\delta^{RB}(z) = [1 - (\sigma^2/(\sigma^2 + \tau^2))](z - \mu) + \mu \tag{12}$$

or

$$\delta^{RB}(z) = z - (c^{0.5}/|z - \mu|)[K_1(w)/K_0(w)](z - \mu) \tag{13}$$

where

σ^2 variance of sample;
τ^2 prior variance;
z sample mean parameter estimate;
μ prior mean parameter estimate;
c $2M\sigma^4/(\sigma^2 + \tau^2)$;
K_0, K_1 modified Bessel functions, with argument w;
w $0.5|z - \mu|c^{0.5}/\sigma^2$.

The choice between the RB estimators in (12) and (13) is based on:

$$|z - \mu|^2 < 2\phi(\sigma^2 + \tau^2) \tag{14}$$

If the inequality holds, then the choice is the Bayes estimator, equation (12); if it does not hold, then the choice is the empiri-

cal Bayes estimator, equation (13). M and ϕ in the above expressions are determined based on the modeler's decisions concerning risk. The decision, stated above, to accept an estimator that may be as much as 20% higher in risk than the maximum likelihood estimator means that $M = 0.20$. The decision to lose only about 20% of the risk gain associated with the Bayes estimator means that [*Berger*, 1982]:

$$0.20 = e^{-\phi}(1 - \phi + M)$$

or

$$\phi = 0.78$$

For the fish response models, three orthogonal components were created (the **U** matrix) using singular value decomposition. OLS regression parameters were estimated and then converted to discriminant function parameters, all in the new orthogonal metric. Berger's RB estimator was applied separately to each discriminant function parameter (a_1, a_2, and a_3) for brook trout and for lake trout. Based on Berger's criterion (equation (14)), two of the parameters are Bayes and one is empirical Bayes for each of the fish species. The EB parameter had Bayes prior and sample distributions with little overlap, implying that the Bayes normal-gamma conjugate family could produce a posterior for this parameter that is not supported by either the prior or sample separately. This provides intuitive justification for the robust Bayes estimator applied to independent parameters.

The Berger RB models are, for brook trout:

$$Pr(P) = 1/[1 + \exp\{3.892 - 2.667pH + 2.298 \ln (Ca)\}] \quad (15)$$

and for lake trout:

$$Pr(P) = 1/[1 + \exp\{5.615 - 2.522pH + 1.905 \ln (Ca)\}] \quad (16)$$

As expected, when these models are applied to predict the probability of presence, the predictions tend to lie between the maximum likelihood predictions and the predictions from a strictly judgment-based model from Dr. Baker's prior alone.

The *p*H and ln (Ca) parameters in (15) and (16), like those in (2) and (3), are of opposite sign, which is counter to scientific theory. Despite the fact that opposite signs do not conform to Dr. Baker's judgment concerning fish response to changes in *p*H and calcium concentration, models for both species fitted to her elicited responses had opposite signs on the *p*H and ln (Ca) parameters [*Reckhow*, 1987]. This result occurs because of the strong correlation between *p*H and ln (Ca) in the data set; the fact that analyses of responses from an expert contradict the expert's view on scientific relationships underscores the dangers in drawing causal conclusions from a statistical analysis.

CROSS VALIDATION

While the robust Bayes models may have a strong intuitive basis, they must also exhibit superior performance in order to justify their use. Here performance means prediction of fish species response to lake acidification in a regional data base. Prediction error for the maximum likelihood (ML) and robust Bayes (RB) models was estimated using FIN. However, prediction error calculated using FIN results in a test biased in favor of the ML models, since the maximum likelihood parameters are estimated from FIN. *Efron* [1986] presents a relationship to estimate this bias; *Reckhow et al.* [1987] applied this to the FIN fish response models.

Recently, two new data sets were acquired which provide the opportunity to evaluate the models on data other than FIN. The Adirondack Lake Survey Corporation (ALSC) collected data in the Adirondacks in 1984 and 1985; these data are not included in FIN. In addition, the Ontario Ministry of Natural Resources (OMNR) surveyed lakes in Ontario [*Beggs et al.*, 1985]. The ALSC data include 202 lakes for which historical presence (pre-1970) of brook trout was confirmed [*Baker et al.*, 1987], while the OMNR data include 192 lakes for which lake trout presence in the past was confirmed.

TABLE 3. Brier Score

Model	Data Set: FIN Brook	FIN Lake	ALSC Brook	Ontario Lake
ML	0.836	0.817	0.799	0.664
RB	0.816	0.805	0.862	0.858

These two data sets were used to compare the maximum likelihood and robust Bayes models. Two statistics were calculated to provide the basis for comparison. Classification error, with $Pr(P) = 0.50$ as the presence/absence cutoff, was selected because the number or proportion of lakes in a region that have lost fish populations is an important quantity to policy makers. A second quantity, the Brier Score B [*Harrell and Lee*, 1985] is

$$B = 1 - \sum (\text{Predicted} - \text{Observed})^2/n \quad (17)$$

which provides a continuous measure of fit, similar to standard error. The Brier Score provides a measure of the average difference between observed and predicted presence/absence; it is expressed so that $B = 1$ represents perfect predictions and $B = 0$ represents perfectly opposite predictions.

Table 3 presents the Brier Score for FIN and for the cross-validation data sets, and Table 4 presents the percent classification error for the same data sets. Probably the most striking feature of these tables is the consistency (robustness) of performance of the robust Bayes (RB) models across the data sets. The RB models are essentially comparable to the maximum likelihood (ML) models on the FIN data set and are far superior to the ML models on the cross-validation data sets. It is encouraging to note that the results of this cross-validation test indicate that the RB models do not exhibit an increase in error when applied to a new data set.

For FIN alone the ML models are better than the RB models on the Brier Score but not on the classification error. This is not surprising since the maximum likelihood logistic regression procedure is not a classification error minimization procedure. The markedly inferior error statistics for the ML models on the cross-validation data sets appears to be due in large part to the response of the ML models at high levels of calcium. For the ML models the predicted probability of pres-

TABLE 4. Percent Classification Error

Model	Data Set: FIN Brook	FIN Lake	ALSC Brook	Ontario Lake
ML	19.6	35.0	32.1	56.3
RB	23.9	20.0	19.9	20.8

ence at high calcium concentrations tends to be lower than the predicted probability of presence for the RB models. Since the ML models are based on behavior exhibited in the data sets, this suggests that lakes high in calcium in FIN are relatively more likely to be absence lakes than are lakes high in calcium in the cross-validation data sets.

CONCLUSIONS

A robust Bayes parameter estimation approach that selects between Bayes and empirical Bayes offers an intuitively appealing option to conventional approaches. The cross-validation test provided strong evidence that the robust Bayes models are able to describe fish response to acid deposition in different regional data sets with consistently lower error statistics than are the maximum likelihood models. This is a robustnesslike quality, but it is to be distinguished from the robustness that is provided by Berger's RB approach. Berger's robust Bayes methods are intended to provide good classical risk performance even if the Bayes prior is misspecified. The robustness of the RB fish response models is due to the fact that the Bayes and empirical Bayes priors adjust for the misspecification of the ML parameters as descriptors of behavior in the cross-validation data sets.

The elicited Bayes prior is given by the expert as a global prior that is specific to a particular fish species but not to a particular region. Thus it should not be surprising that use of this prior may result in a Bayes or robust Bayes model that can be applied with success across regional data sets. Given the global nature of the Bayes prior, a Bayesian analysis might be considered appropriate here. However, at this point, the robust Bayes approach is preferred. When the prior and sample information are in general agreement, the RB estimator will equal the mean of the Bayes posterior distribution (variance known), but when prior and sample disagree (as they might when the prior is obtained from experts unfamiliar with the elicitation of prior probabilities), the robustness of the estimator protects against large risk. If the experts become well calibrated, then a Bayesian analysis should be chosen.

The results of this study are viewed as support for robust Bayes estimators in models of fish response to acidification. Given the small size of the FIN data set, future work on this topic should include maximum likelihood parameter estimation with the ALSC and OMNR (cross-validation) data sets, and use of these ML estimates in new robust Bayes models. Other information, such as experimental studies, could also be used to construct the Bayes or robust Bayes prior. Bayesian and robust Bayes methods offer the opportunity to go beyond classical analyses and use the valuable collateral information for the improvement of predictive models.

Acknowledgments. Joan Baker was extremely helpful in many discussions of fish response to water chemistry. Useful comments were also provided by Mitch Small and by the Institute of Statistics and Decision Sciences seminar group at Duke University. Data were provided by the Adirondack Lake Survey Corporation and the Ontario Ministry of Natural Resources. Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency through CR-812641-01-0 to K. H. Reckhow, it has not been subjected to Agency review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

REFERENCES

Baker, J. P., State of the science: Effects of acidification on aquatic biota, North Carolina State University acid deposition program report to U.S. Environmental Protection Agency, Environ. Prot. Agency, Corvallis, Oreg., 1985.

Baker, J. P., and T. B. Harvey, Critique of acid lakes and fish population status in the Adirondack region of New York State, final report to U.S. Environmental Protection Agency, Environ. Prot. Agency, Corvallis, Oreg., 1984.

Baker, J. P., and C. Schofield, Aluminum toxicity to fish in acidic waters, *Water Air Soil Pollut., 18*, 289–309, 1982.

Baker, J. P., C. S. Creager, and S. W. Christensen, Identification of critical values for effects of acidification on fish populations, draft report to the U.S. Environmental Protection Agency, Environ. Prot. Agency, Washington, D.C., 1987.

Beggs, G. L., J. M. Gunn, and C. H. Olver, The sensitivity of Ontario lake trout (Salvelinus namaycush) and lake trout lakes to acidification, *Ont. Fish. Tech. Rep. Ser. 17*, Ont. Minist. of Nat. Resour., Toronto, Ont., Canada, 1985.

Berger, J. O., Bayesian robustness and the Stein effect, *J. Am. Stat. Assoc., 77*, 358–368, 1982.

Berger, J. O., The robust Bayesian viewpoint, in *Robustness of Bayesian Analyses*, edited by J. Kadane, pp. 63–124, North-Holland, Amsterdam, 1984.

Brown, D. J. A., The influence of calcium on the survival of eggs and fry of brown trout at *p*H 4.5, *Bull. Environ. Contam. Toxicol., 28*, 664–668, 1982.

Brown, D. J. A., and K. Sadler, The chemistry and fishery status of acid lakes in Norway and their relationship to European sulphur emissions, *J. Appl. Ecol., 18*, 433–441, 1981.

Efron, B., The efficiency of logistic regression compared to normal discriminant analysis, *J. Am. Stat. Assoc., 70*, 892–898, 1975.

Efron, B., How biased is the apparent error rate of a prediction rule?, *J. Am. Stat. Assoc., 81*, 461–470, 1986.

Efron, B., and C. Morris, Limiting the risk of Bayes and empirical Bayes estimators, Part 1, The Bayes case, *J. Am. Stat. Assoc., 66*, 807–815, 1971.

Efron, B., and C. Morris, Limiting the risk of Bayes and empirical Bayes estimators, Part 1, The empirical Bayes case, *J. Am. Stat. Assoc., 67*, 130–139, 1972.

Green, P. E., *Analyzing Multivariate Data*, 519 pp., Dryden Press, Hinsdale, Ill., 1978.

Halperin, M., W. C. Blackwelder, and J. I. Verter, Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches, *J. Chronic Dis., 24*, 125–158, 1971.

Harrell, F. E., and K. L. Lee, A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality, in *Biostatistics: Statistics in Biomedical, Public Health, and Environmental Sciences*, edited by P. K. Sen, pp. 167–175, North-Holland, Amsterdam, 1985.

Kahneman, D., P. Slovic, and A. Tversky, *Judgment Under Uncertainty: Heuristics and Biases*, 433 pp., Cambridge University Press, New York, 1982.

Kuczera, G., Combining site specific and regional information: An empirical Bayes approach, *Water Resour. Res., 18*, 306–314, 1982.

Landwehr, J. M., N. C. Matalas, and J. R. Wallis, Note on the applicability of the James-Stein estimator in regional hydrologic studies, *Water Resour. Res., 20*, 1630–1638, 1984.

Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, 401 pp., Cambridge University Press, New York, 1983.

Mandel, J., Use of singular value decomposition in regression analysis, *Am. Stat. 36*, 15–24, 1982.

Marin, C., Parameter estimation in water resources planning and management: Optimal actions or optimal parameters?, *Water Resour. Res., 22*, 353–360, 1986.

Morris, C., Parametric empirical Bayes inference: Theory and applications, *J. Am. Stat. Assoc., 78*, 47–65, 1983.

Muniz, I. P., H. M. Seip, and I. H. Sevaldrud, Relationship between fish populations and *p*H for lakes in southernmost Norway, *Water Air Soil Pollut., 23*, 97–113, 1984.

Pindyck, R. S., and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, 630 pp., McGraw-Hill, New York, 1981.

Press, S. J., and S. Wilson, Choosing between logistic regression and discriminant analysis, *J. Am. Stat. Assoc., 73*, 699–705, 1978.

Reckhow, K. H., Robust Bayes models of fish response to lake acidification, in *Systems Analysis in Water Quality Management*, edited by M. B. Beck, pp. 61–72, Pergamon, Elmsford, N.Y., 1987.

Reckhow, K. H., R. W. Black, T. B. Stockton, Jr., J. D. Vogt, and J. G. Wood, Empirical models of fish response to lake acidification,

253 pp., report to U.S. Environmental Protection Agency, Environ. Prot. Agency, Corvallis, Oreg., 1985.

Reckhow, K. H., R. W. Black, T. B. Stockton, Jr., J. D. Vogt, and J. G. Wood, Empirical models of fish response to lake acidification, *Can. J. Fish. Aquat. Sci.*, *44*, 1432–1442, 1987.

Rolph, J. E., A. P. Williams, and C. L. Lee, Medical school admission and residence of applicants: Empirical Bayes estimates of logit coefficients, *J. Educ. Stat.*, *4*, 291–323, 1979.

Sevaldrud, I. H., and I. P. Muniz, Acid lakes and inland fisheries in Norway: Results of interview surveys 1974–1979, *SNSF-Proj. IR77/80*, Norw. Inst. for Water Res., Oslo, Norway, 1980.

Winkler, R. L., Prior distributions and model-building in regression analysis, in *New Developments in the Applications of Bayesian Methods*, edited by A. Aykac and C. Brumat, pp. 233–242, North-Holland, Amsterdam, 1977.

Winkler, R. L., W. S. Smith, and R. B. Kulkarni, Adaptive forecasting models based on predictive distributions, Manage. Sci., *24*, 977–986, 1978.

Wright, R. F., and E. Snekvik, Chemistry and fish populations in 700 lakes in southernmost Norway, *SNSF-Proj. TN 37/77*, Norw. Inst. for Water Res., Oslo, Norway, 1977.

Zison, S. W., W. B. Mills, D. Diemer, and C. W. Chen, Rates, constants, and kinetic formulations in surface water modeling, *Rep. U.S. EPA EPA-600/3-78-105*, Environ. Prot. Agency, Athens, Ga., 1978.

K. H. Reckhow, School of Forestry and Environmental Studies, Duke University, Durham, NC 27706.

(Received June 22, 1987;
revised March 16, 1988;
accepted March 17, 1988.)